A PROCEDURE FOR AUTOMATIC DATA EDITING

R. J. Freund and H. O. Hartley, Texas A & M University

1. INTRODUCTION

A cursory study of some presently used editing procedures reveals that a great variety of these procedures have been developed. Some of these are based on simple 'one item at a time' consistency checks and imputations using either common sense consistency principles, 'historical data', 'hot-deck' or 'regression estimates'. Others employ complex networks of interlocking checks and associated imputations. This great variety of editing procedures reflects the impact of two often conflicting desiderata:

(a) An effective editing procedure must recognize the particular error patterns as well as inconsistency-and-error correlations arising in a particular survey.

(b) An effective editing procedure must have a comparatively simple logic and must be easy to program as otherwise there is a tendency not to use automatic editing at all.

Desideratum (a) calls for a procedure 'custom made' for the particular survey and thus uses a logic specifically oriented to a particular study; this requires considerable programming effort for each survey. Desideratum (b) calls for a procedure which is easily understood and applied by the programmer, but thereby tends to bypass a detailed scrutiny of the specific error patterns which the specialist in a particular survey area would call for.

In attempting to reconcile (a) and (b) we have tried to make contributions on the following lines:

(a) We have attempted to develop a standard editing procedure which is to be implemented by some simple macro-codes or special forms, and

(b) we have developed a procedure for producing relatively consistent data when several restrictions must be met.

The general procedure consists of:

1. The Gross Check.

2. The Internal Consistency Check.

If the internal consistency check fails, we will correct the data by

3. The Least Squares Correcting Procedure.

1.1 THE GROSS CHECK

Each item of the input record is checked for gross errors. Such errors are characterized by the data being completely unrealistic or out of line. Data items found to violate such gross error checks are imputed, one at a time, by relatively simple and straightforward imputations such as the use of historical data, 'hot decks', the use of ratios known to be usually consistent, corrections for misplacement of decimals or incorrect units of measurement, etc. A record will be kept of all data items for which imputations have been made in this editing phase.

1.2 THE INTERNAL CONSISTENCY CHECK

The data will be checked to see if certain internal consistencies are satisfied within a sufficient tolerance. Thus, for example, acreages in various crops must add up to total acreage in crops, quantity times price must be equal to value, etc. Any time a check fails to be fulfilled, attempts may be made to impute data by some of the same procedures as were used in the gross check. During this phase of the editing procedure, a notation will be made of all data items involved in checks which are not satisfied.

1.3 THE LEAST SQUARES CORRECTING PROCEDURE

A least squares correcting procedure will be used if simple, one-at-a-time correcting procedures do not produce data which will satisfy all internal consistency checks. It is proposed that this procedure will substitute for the complex interlocking networks of checks and imputations which are usually used. It is hoped that the use of this procedure will not occur too frequently since it will take a moderate amount of computer time. However, it will require no complex, custom-made networks and will, therefore, be generally useful for many situations.

This procedure also makes use of the fact, noted above, that some items in the data are more 'suspect' than others due to their either being subjected to corrections in the gross check or being involved in some of the consistency checks which failed to satisfy tolerances. This will be done by the use of weighted least squares, where the weights will indicate the reliability of the data. These weights can, of course, also be used to indicate the 'usual' reliability of individual data points, recognizing that some data are usually more reliable than others.

1.3.1 General Considerations

We start with the previously stated premise that we want to satisfy (to some degree) certain (linear) consistency equations by 'correcting' some of the input variables. Some of the consistency equations are more important than others, i.e., some must be satisfied to a greater degree of accuracy than others. Likewise, some input data are assumed to be more reliable than others and consequently some data should be changed less than others by the correcting procedure, but the corrected data should be as 'close' as possible to the original data. This is to be accomplished by minimizing the weighted sum of squares of the discrepancies of the consistency equations plus the weighted sum of squared differences between original and corrected data; the weights are used to indicate importance of restrictions and/or reliability of data.

The equation for the sums of squares to be minimized is:

$$SSC = \sum_{j}^{n} w_j(x_j-y_j)^2 + \sum_{i}^{n} u_i\left(\sum_{j}^{n} a_{ij}x_j\right)^2$$

where

x_j is the j-th corrected datum

y_j is the j-th original datum

w_j is the weight given to the i-th datum, a large weight indicates more reliable data,

$\sum_{j}^{n} a_{ij}x_j$ = the j-th consistency equation of the form $\sum_{j}^{n} a_{ij}x_j = 0$,

u_j is the weight of the j-th consistency, a large weight indicates an important consistency.

The minimization is with respect to the x_j;

thus the corrected data is chosen to minimize this sum of squares.

The minimization is accomplished as follows:

$$\frac{\partial SSC}{\partial x_k} = 2w_k(x_k-y_k) + \sum_{i}^{m} u_i \left(2\sum_{j}^{n} a_{ij}x_j a_{ik}\right) = 0,$$

$$k = 1, 2, \ldots, n.$$

Rearranging terms we have:

$$w_k(x_k-y_k) + \sum_i u_i a_{ik} \sum_j a_{ij}x_j = 0, \quad k = 1, 2, \ldots, n.$$

In matrix form this can be written:

$$D_w(\underline{x}-\underline{y}) + A'D_uA\underline{x} = 0$$

where

D_w is a (n x n) diagonal matrix of the w weights (for data),

$\underline{y}$ is a (n x 1) vector of original data,

$\underline{x}$ is a (n x 1) vector of corrected data,

A is the matrix of coefficients of the restrictions, hence these can be written $A\underline{x} = \underline{0}$, where $\underline{0}$ is a vector of zeroes,

D_u is a (m x m) diagonal matrix of the u weights (for the restrictions).

Solving for $\underline{x}$ we have

$$\underline{x} = (D_w + A'D_uA)^{-1}D_w\underline{y}$$

This expression is, of course, easily solved by high speed computers. Note that A will be usually predetermined for an entire study; D_u may also be constant for an entire study and hence $A'D_uA$ need be computed only once.

1.3.2 The Determination of Weights

The weights of the data points should exhibit

(a) The basic variability of the datum,
(b) the 'usual' reliability with which the datum is reported, and
(c) the reliability of the datum in a specific record.

(a) Basic Variability

Data which is basically variable is more subject to corrections and should, therefore, receive smaller weights. This type of variability is often associated with size of unit of measurement. Thus small items should receive smaller corrections and large items large corrections; data which should be zero should ideally receive no correction.

(b) 'Usual' Reliability

Some data are naturally recorded more accurately than others. For example, tobacco acreage are very precisely known due to strict acreage controls whereas woodland acreages may not be well known, particularly if woodlands are used partially as pastureland.

(c) Reliability of a Specific Record

It is assumed that the record which is subjected to a least squares correction has failed in an initial, relatively simple, editing-correcting sequence (see above). If a data point has been subject to a gross error correction or has been involved in several unsatisfied consistency checks, it is most likely in error. Thus, weights of items involved in a gross error correction or non-satisfied consistency check will be reduced in proportion to the number of involvements.

The weights for the consistency equations (u_i)

should exhibit the importance of the consistency equations, i.e., the degree in which the equation must be satisfied. Sometimes it is vital for the purposes of the study that a certain consistency check is accurately satisfied while other checks are not as critical. The use of (relatively) large weights for some equations will assure small discrepancies in these equations.

It should be noted that weights are relative and it is the ratio of large to small weights which is of importance. The magnitude of this ratio for practical use is subject to further study; initial experimentation indicates ratios of 5/1 to 20/1 are needed for effective control, i.e., differentiation of magnitude of corrections. It can be further noted here that we are attempting to correct data points to conform to certain consistency checks. Thus it is reasonable that weights for data should be smaller than weights for restrictions. On the other hand, the most reliable individual data points should probably not be corrected. Thus a procedure for assigning weights should assign nearly equal and relatively large weights for the most reliable data points and most important consistency equations; the least reliable equations should have weights possibly 1/5 as large and the least reliable data points 1/20 to 1/100 as large.

The entire procedure outlined above can be summarized in a flow chart as outlined in Figure 1.

2. EXAMPLE

We will use as examples the data from some hypothetical farms, using selected information as recorded in the Bureau of Census, Farm Questionnaire Sample Survey of Agriculture, 1961 (Form No. 60-02-548.4). Fourteen items involving acreages have been selected for use (see Table 1).

The "w weights" indicating the 'basic variability' and 'usual reliability' of the data are given in Table 1. Thus, for example, acres owned and acres of cropland should be reliably recorded since exact knowledge of these is required for taxes and government programs; the rental - sub rental acreages can be considered confusing, and thus, are more likely to be in error.

The restrictions are given in Table 1 as coefficients (reading vertically) of equations that should equal zero. Thus, restriction 5 states that total acres in place is equal to acres owned and not rented-out plus acres rented but not sub-rented. The "u weights" given at the bottom indicate the importance of the restriction; thus it is considered important that acres in place agree both with respect to rental arrangement and land use (restriction 2, u = 85) but breakdown of land use is open to questions of unaccounted land (fences, roads, etc.) and double cropping (restriction 1, u = 10).

2.1 GENERATION OF DATA

We shall attempt to evaluate the above outlined editing procedures by their use on some artificially generated 'incorrect' data. We assume that we have a large number of identical schedules into which we introduce random errors. It is then relatively easy to see how close to the "correct" data the editing procedures actually come.

There are essentially two decisions to be

Table 1

Data For Example

DATA					RESTRICTIONS (coefficients, read down)						
Variables Acres	Item No. from Schedule	Correct Acres	Mnemonic	"w" weight	1 Land Use	2 Tot= Tot	3 Land Use	4 Rent Out	5 All Tenure	6 Rent In	7 Total Acres
"u" weight →					10	85	80	40	35	40	50
Owned	8A	160	OWN	10				+1			
Rented to	8B	10	RENTO	6				-1			
Not rent to	8C	150	RENTONO	1				-1	-1		-1
Rent from	9A	65	RENTFM	3					-1	+1	
Subrent	9B	0	SRENT	6					+1	-1	
Not Subrent	9C	65	SRENTNO	1						-1	-1
Total in place*	10	215	TIP	6		-1			+1		+1
Cropland total	61	180	CROPT	10	+1		-1				
Pasture	61A	60	PAST	3	-1						
Gov't Program	61B	90	GOVTP	10	-1						
Other	61C	20	FFETC	3	-1						
Harvested	61D	10	HARV	5	-1						
Total in place*	66	215	TAIP	6		+1	+1				
Other uses	(62+63+ 64+65)	35	OWSES	1			-1				
Tolerated Limits					3%	2%	2%	2%	3%	2%	2%

*This is requested twice in schedule.

in the generation of data with errors:

1. Whether a particular item be correct or incorrect.
2. If a particular item is incorrect, what type of error it should exhibit.

The procedure generated errors in two steps. First, a random number was generated to correspond to each item in the schedule and if the random number was less than $1/1.5w$, then the item was designated as being incorrect. This procedure does, of course, generate a much larger than usual number of errors (the most reliable items had W = 10, hence over 6% of even these are in error), but we do not wish to waste computer time not correcting many "good" records.

Once an item was designated as being in error a second random number was generated and the type of error was assigned as indicated in Table 2. Thus, for example, if the random number was between 0 and .1 it is assumed

Table 2

TYPES OF ERRORS TO BE GENERATED

Random Number	Type of Error
0 —> .1	Blank
.1 —> .2	+ 10%
.2 —> .3	- 10%
.3 —> .4	+ 5%
.4 —> .5	+ 5%
.5 —> .6	- 5%
.6 —> .7	- 5%
.7 —> .8	* 10
.8 —> .9	* 0.1
.9 -> 1.0	Return

there was a blank in that particular portion of the schedule, a random number between .7 and .8 would indicate that there had been a scaling error of a factor of 10. The random number between .9 and 1.0 was not used for a specific error and hence, if this occurred, a second random number would be generated and the error assigned according to the second random number. This allows one additional type of error to be introduced if it is desired to do so.

2.2 RESULTS

Several different sampling experiments were performed using minor variations of the above outlined editing procedure. In all there were eleven sampling experiments. Since the random number generator always has the same starting value, all samples are based on identical sets of "incorrect" data, a direct comparison of the results is quite meaningful. These comparisons are afforded by the summary in Table 3 which shows, for the originally generated data and the results of various editing procedures, the mean difference between each of the correct and corrected items and the standard deviation of the items. At the bottom of the Table are the sums of absolute mean deviations and the standard deviations, also the sum of squared deviations and sum of variances. It should be noted that these standard deviations are based on the sum of squared differences of individually corrected items and the sample means of the corrected items rather than the sum of squared deviations from the true values of the items. This latter figure can, of course, be generated, but would not be much larger than those given.

Of course, any editing procedure which resulted in a mean difference of 0 and a standard deviation of 0 would be ideal; needless to say this has not been realized and is not likely to be realized. It is difficult to make a value judgment as to whether it is more important that the mean difference be 0 or that the variance be small; since most data of this type is used for summaries, it is most likely more important that the mean differences be small. It should be remembered that the results of the least squares procedure do indeed guarantee that we have consistent results; that is, the various consistency checks will be almost completely satisfied.

In Table 3 the first column provides the correct values and the second set of two columns provides the mean differences and standard deviations of the data as generated by the random error generator. It can be seen that this procedure was quite successful in generating "incorrect" data. There appears to be a definite upward bias in practically all the items; this is due to the fact that we had a ten per cent chance of generating an item ten times too large. The second set of two columns indicates what the data would look like if it were subjected only to the gross check and imputations procedure (item 2 on flow chart). It can be seen that this procedure does definitely improve the quality of the data and it might well be argued that one should stop there. It should be noted, however, that there is no guarantee that the results will be consistent for any given schedule or, indeed, for averages derived from this data.

The next four sets of two columns are the result of the editing procedure as outlined in Figure 1 except that no corrections were made in the consistency check stage (item 5 on flow chart is bypassed). In other words, the consistency equations were checked and if any one equation was not satisfied the least squares procedures was used. All checks were, however, made in order to provide the count of the number of times items are involved in unsatisfied consistency checks.

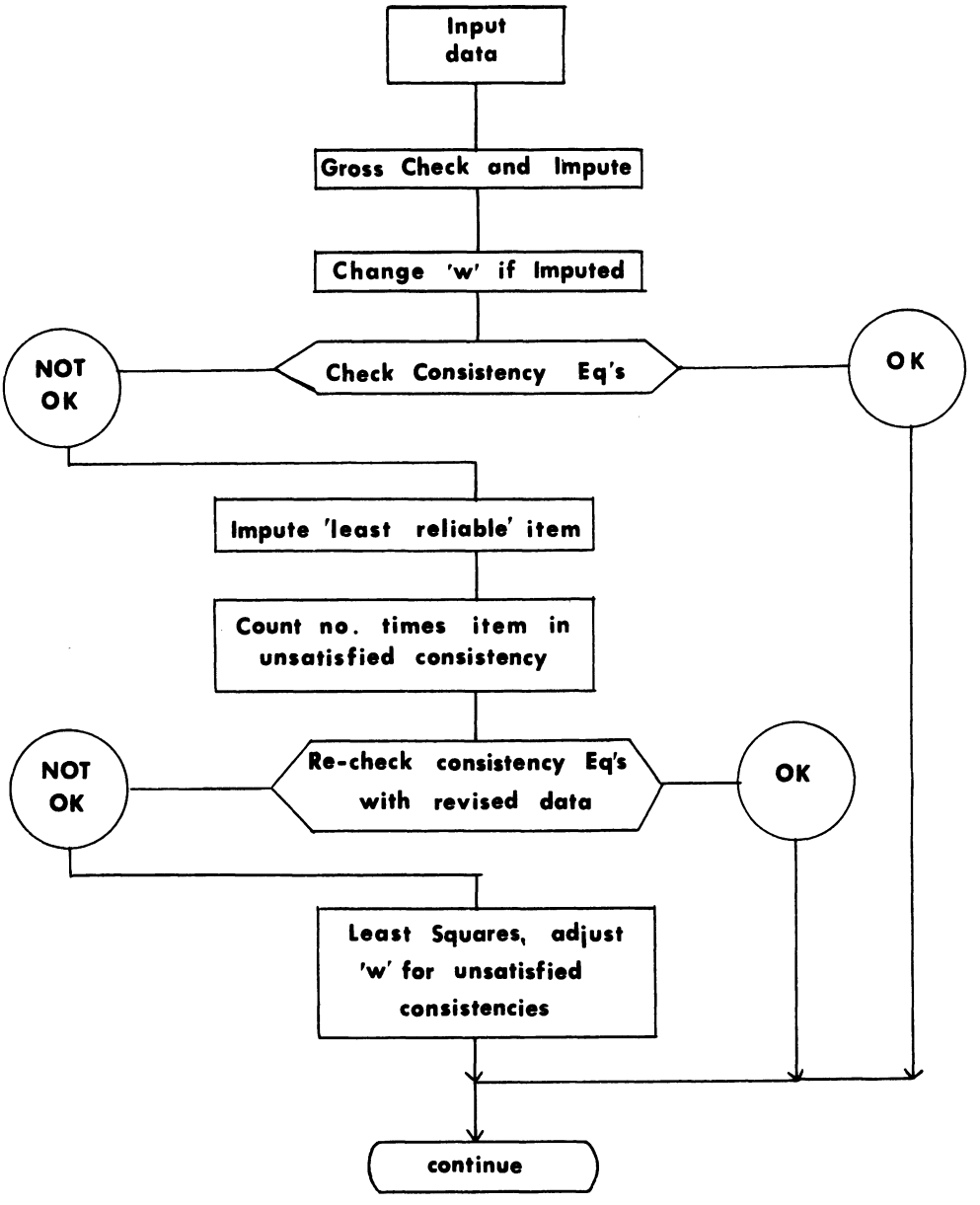
The reasoning behind the elimination of this particular step in the editing procedure is that

Table 3: Results of Sampling Experiment

						Gross Check and Least Squares Correction								Gross Check, Consist. Corr. and L. S.				Gross Check and L. S. all w's start at 10 for L. S.			
		Generated		Gross Check		Original Wts.		Wts. adj. for Gr. Ch.		Wts. fully Adj. WAF#1		Wts. fully Adj. WAF#2		Wts. fully Adj. WAF#1		Wts. fully Adj. WAF#2		Wts. fully Adj. WAF#1		Wts. fully Adj. WAF#2	
ITEM	Correct μ	diff	σ	diff	σ	diff	σ	diff	σ	diff	σ	diff	σ	diff	σ	diff	σ	diff	σ	diff	σ
OWN	160	7.5	126.5	-4.0	23.7	-3.5	20.7	-3.5	20.7	-3.1	19.2	-3.0	19.2	-3.3	20.2	-3.3	20.2	-3.1	19.2	-3.1	19.3
RENTO	10	.2	6.5	.2	6.5	- .5	8.4	- .5	8.2	-1.2	9.0	-1.3	9.4	-1.0	7.8	-1.0	7.9	-1.2	9.0	-1.3	9.4
RENTNO	150	96.5	365.4	-8.8	35.4	-2.8	16.3	-2.9	16.8	-1.8	13.6	-1.7	14.1	-2.3	15.4	-2.2	15.3	-1.8	13.6	-1.7	14.1
RENTFM	65	8.0	77.9	-2.5	12.0	- .7	10.1	- .6	10.1	- .6	9.0	- .5	9.5	- .3	9.5	- .3	9.6	- .6	9.0	- .5	9.5
SRENT	0	0.0	0.0	0.0	0.0	- .9	4.3	- .9	4.3	- .9	4.5	-1.0	4.9	-1.1	4.5	-1.1	4.7	- .9	4.5	-1.0	4.9
SRENTNO	65	37.2	154.3	-5.5	17.4	+ .1	11.7	+ .2	12.0	+ .4	10.7	.5	11.8	+ .8	11.3	+ .8	11.6	.4	10.7	+ .5	11.8
TIP	215	23.1	215.9	-1.1	14.7	-3.6	13.9	-3.6	13.7	-1.4	11.2	-1.1	11.4	-1.5	12.7	-1.4	13.3	-1.4	11.2	-1.2	11.4
CROPT	180	18.5	181.3	-3.0	22.9	-2.4	20.3	-2.4	19.7	-2.2	17.5	-1.7	14.7	2.1	17.8	-1.6	15.2	-2.2	17.5	-1.7	14.7
PAST	60	11.7	85.3	-2.1	10.9	-3.0	12.2	-3.1	12.8	-3.1	13.0	-2.9	12.4	-1.5	13.1	-1.3	12.5	-3.1	13.0	-2.9	12.4
GOVTP	90	- .5	6.0	- .1	8.3	-1.2	8.5	-1.1	8.6	-1.2	8.6	-1.1	8.5	-1.4	8.8	-1.3	8.7	-1.1	8.6	-1.1	8.5
FFETC	20	3.2	25.4	3.2	25.4	+2.3	19.2	+2.4	19.7	+2.4	18.9	+2.6	18.8	+1.6	18.8	1.8	18.7	2.4	18.9	+2.6	18.8
HARV	10	+ .4	6.4	+ .4	6.4	- .2	7.9	- .2	8.2	- .2	8.4	- .1	8.0	- .7	8.1	- .6	7.8	- .2	8.4	- .1	8.0
TAIP	215	12.8	168.5	-1.7	20.0	-3.5	14.0	-3.6	13.8	-1.5	11.2	-1.1	11.4	-1.5	12.7	-1.4	13.3	-1.4	11.2	-1.1	11.4
OUSES	35	21.0	84.4	-2.9	9.6	- .1	17.4	- .2	17.3	+ .7	14.1	+ .5	10.3	+ .7	14.3	.4	10.9	.7	14.1	+ .5	10.3
Sum (abs)		240.6	1503.8	35.5	213.2	24.8	184.9	25.2	185.9	20.7	168.9	19.1	164.4	19.8	174.0	18.5	169.7	20.5	168.9	19.3	164.5
Sum Sq.		12,444.62		166.31		68.40		70.46		42.01		37.03		35.78		32.29		41.49		37.87	
		302,434.88		44,410.34		2,778.33		2,801.11		2,277.01		2,134.86		2,443.64		2,299.45		2,277.01		2,138.71	

Figure 1

Flow Chart of Editing Procedure



the consistency-correcting procedure was quite arbitrary in that if a particular consistency equation failed, the item in that equation with the lowest "w" weight was imputed by subtraction. It is obvious that, in many cases, a correct item may be altered by this procedure, thus creating other inconsistencies which will cause further "corrections" of correct data. Thus this procedure will often create incorrect data which is nevertheless consistent and will thus not be further edited.

In the first two columns of this set (entitled Original Weights) the "w" weights are never adjusted (either in the gross check or in the consistency check) and thus the least squares is entered with the "w" weights as originally indicated in Table 1. In the second set of columns the "w" weights are only adjusted in the gross check phase, where "w" weights are divided by 4 for any item which is imputed at that stage.

The third and fourth sets of columns in this group are results of the procedure (as above) with two different sets of factors used to adjust the "w" weights for the number of involvements in unsatisfied consistency checks. The third set of columns corresponds to the weight adjustment factors as given in Table 4; this Table indicates the divisor for the "w" as a function of the possible number involvements in consistency checks and the actual number of involvements in unsatisfied consistency checks. The fourth set of columns is a result of the use of the weight adjustment factors in Table 5; these adjustments are more "severe" and actually increase "w" weights in case an item is often involved in satisfied consistency checks.

Table 4
Weight Adjustment Factors No. 1

No. of Actual Involvements	No. of Possible Involvements 1	2	3
0	1	1	1
1	5	2	1
2		10	5
3			20

Table 5
Weight Adjustment Factors No. 2

No. of Actual Involvements	No. of Possible Involvements 1	2	3
0	1	1/5	1/10
1	2	1/2	1/5
2		10	2
3			50

In the next two sets of columns the entire editing procedure as given in Figure 1 is used with the two different weight adjustment factors (i.e., Tables 4 and 5, respectively).

The last two sets of columns are intended to show what happens when incorrect "w" weights are used. In this particular sampling experiment the "w" weights as presented in Table 1 are used to generate the data, but for the least squares procedure all weights are initially set equal to ten before being adjusted in the gross check and consistency phases as before with the two different sets of weight adjustment factors. This set of sampling experiments was performed in order to see if it is really very important to initially assign weights indicating prior knowledge of the reliability of items.

The results of Table 3 can be summarized as follows:

a. The gross check does improve the data to a great extent, but the use of the least squares procedure definitely improves the data even further.

b. It appears that the adjustment of weights in the gross check phase is not of much help, but that the adjustment of weights from the consistency equations is useful.

c. From this point on there is not much difference in the results among the procedures and at present it would seem that the elimination of the consistency imputations and the use of weight adjustment factor 2 without prior assignment of differentiated weights is the optimum procedure.

More work of this type is certainly desirable before more definite conclusions can be drawn.